UNIVERSITY OF TECHNOLOGY SYDNEY
Faculty of Engineering and Information Technology

Small Object Detection and Recognition Using Context and Representation Learning

by

Yue Xi

A Thesis Submitted
in Partial Fulfillment of the
Requirements for the Degree

Doctor of Philosophy

Sydney, Australia

2021

Certificate of Authorship/Originality

Yue Xi declares that this thesis, is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the Faculty of Engineering and Information Technology at the University of Technology Sydney.

I certify that this thesis has been written by me. Any help that I have received in my research and in the preparation of the thesis itself has been fully acknowledged. In addition, I certify that all information sources and literature used are quoted in the thesis.

This thesis is from the candidate's research conducted for the University of Technology Sydney and Northwestern Polytechnical University Collaborative Doctoral Degree.

I also certify that this thesis has not been submitted for qualifications at any other academic institution. This research is supported by the Australian Government Research Training Program.

Production Note:
Signature: Signature removed prior to publication.

Date: 26 January 2021

ABSTRACT

Small Object Detection and Recognition Using Context and Representation Learning

by

Yue Xi

Small object detection and recognition is very common in real world applications, such as remote sensing images analysis for Earth Vision, Unmanned Aerial Vehicle vision and video surveillance for identity recognition. Recently, the existing methods have achieved impressive results on large and medium objects. But the detection and recognition performance for small or even tiny objects is still far from satisfaction.

The problem is highly challenging because small objects in low-resolution images may contain fewer than a hundred pixels, and lack sufficient details. Context plays an important role on small object detection and recognition. Aiming to boost the detection performance, we propose a novel discriminative learning and graph-cut framework to exploit the semantic information between targeting objects' neighbours. What is more, to depict a local neighbourhood relationship, we introduce a pairwise constraint into a tiny face detector to improve the detection accuracy. At last, to describe such a constraint, we convert the problem of regression that estimates the similarity between different candidates into a classification problem that produces the score of classification for each pair of candidates.

In representation learning, we propose an RL-GAN architecture, which enhances the discriminability of the low-resolution (LR) image representation, resulting in comparable classification performance with that conducted on high-resolution (HR) images. In addition, we propose a method based on a Residual Representation to generate a more effective representation of LR images. The Residual Representation is adapted to fuel back the lost details in the representation space of LR images.

At last, we produce a new dataset WIDER-SHIP, which provides paired images of multiple resolutions of ships in satellite images and can be used to evaluate not only LR image classification but also LR object recognition.

In the domain of a small sample training, we explore a novel data augmentation framework, which extends a training set to achieve a better coverage of varying orientations of objects in a testing data, so as to improve the performance of CNNs for object detection. Then, we design a principal-axis orientation descriptor based on super-pixel segmentation to represent the orientation of an object in an image. We propose a similarity measure method of two datasets based on a principal-axis orientation distribution. We evaluate the performance and show the effectivity of CNNs for object detection with and without rotating images in the testing set.

Dissertation is directed by Professor Xiangjian He and Doctor Wenjing Jia of University of Technology Sydney, Australia, and Professor Jiangbin Zheng of Northwestern Polytechnical University, China.

Dedication

I would like to dedicate this thesis to my wife, Fuxi Ji. I want to thank her for being so patient with me and supportive through these years of my study for the Ph.D. degree. I would also like to thank my parents, Junhou Xi and Lijun Zhang, and my new baby Jirui Xi. All the best wishes to you.

Acknowledgements

Firstly, I would like to express my sincere gratitude to my advisor Professor Xiangjian He for the continuous support of my Ph.D study and the related research, for his patience, motivation, and immense knowledge. His guidance helped me in all time of the research and writing of this thesis.

Besides my supervisor, I would like to thank my co-supervisor Doctor Wenjing Jia for their insightful comments and encouragement, but also for the hard question which incented me to widen my research from various perspectives.

I would also like to thank Professor Jiangbin Zheng, my supervisor in my home university - Northwestern Polytechnical University, for the collaborative PhD degree between UTS and NPU. Thank all students in the Computer Vision and Pattern Recognition Laboratory of UTS for having given me pleasant memories.

Yue Xi

Sydney, Australia, July 2020.

List of Publications

Chapter 3

C-1. **Yue Xi**, Jiangbin Zheng, Xiangjian He, Wenjing Jia and Hanhui Li, "Beyond Context: Exploring Semantic Similarity for Tiny Face Detection." 2018 25th IEEE International Conference on Image Processing (ICIP), PP. 1907-1911, 2018.

J-1. **Yue Xi**, Jiangbin Zheng, Xiangjian He, Wenjing Jia, Hanhui Li, Yefan Xi, Mingchen Feng and Xiuxiu Li, "Beyond context: Exploring semantic similarity for small object detection in crowded scenes." Pattern Recognition Letters pp. 53-60, 137, 2020.

Chapter 4

J-2. **Yue Xi**, Jiangbin Zheng, Wenjing Jia, Xiangjian He, Hanhui Li, Zhuqiang Ren and Kin-Man Lam. "See Clearly in the Distance: Representation Learning GAN for Low Resolution Object Recognition." IEEE Access, 8, 53203-53214, 2020.

Chapter 5

J-4. **Yue Xi**, Wenjing Jia, Jiangbin Zheng, Xiaochen Fan, Yefan Xie, Jinchang Ren and Xiangjian He. "DRL-GAN: Dual-Stream Representation Learning GAN for Low-Resolution Image Classification in UAV Applications." IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, pp. 1705–1716, 14, 2021.

Chapter 6

J-5. **Yue Xi**, Jiangbin Zheng, Xiuxiu Li, Xinying Xu, Jinchang Ren and Gang Xie, "SR-POD: sample rotation based on principal-axis orientation distribution for data augmentation in deep object detection." Cognitive Systems Research, pp. 144–154, 52, 2018. Elsevier.

C-2. Jiangbin Zheng, **Yue Xi**, Mingchen Feng, Xiuxiu Li, Na Li, "Object detection based on BING in optical remote sensing images," 2016 9th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), pp. 504-509, 2016.

Contents

Certificate ii

Abstract iii

Dedication v

Acknowledgments vi

List of Publications vii

List of Figures xiv

List of Tables xix

Abbreviation xxi

1 Introduction 1

1.1 Small Object Detection and Recognition 1

1.1.1 SODR System 3

1.1.2 Image Acquisition 3

1.1.3 Small Object Detection 3

1.1.4 RoIs Processing 4

1.1.5 Small Object Recognition 4

1.2 Key Challenges of SODR and Core Scientific Problem 5

1.2.1 Limited Information for Identifying 5

1.2.2 Limitation of Camera Resolution 5

1.2.3 More Possibilities for Locations of Small Objects and Limited Prior Knowledge 6

1.2.4 Core Scientific Problems 6
1.3 Contributions in the Thesis 7
1.3.1 Overview of the Research in the Thesis 7
1.3.2 Context-based Small Object Detection 8
1.3.3 Representation Learning for Small Object Recognition 9
1.3.4 Principal-axis Orientation Distribution for Data Augmentation 10
1.4 Organisation of the Thesis 10

2 Literature Review and Mathematical Background 13

2.1 Small Object Detection 13
2.1.1 Small and Crowded Object Detection 13
2.1.2 Context in Object Detection 15
2.1.3 Discriminative Learning for Semantic Similarity 16
2.2 Small Object Recognition 16
2.2.1 LR Face or Activity Recognition 17
2.2.2 Hallucination based LR Approaches 17
2.2.3 Representation-transforming based Approaches 18
2.2.4 Spectral Bias in Deep Learning 19
2.3 Data Augmentation for SODR 20

3 Context-Based Small Object Detection 22

3.1 Overview of the Proposed Method 25
3.1.1 Tiny Face Detection 25
3.1.2 Small Object Detection in Remote Sensing Images 26
3.2 Discriminative Learning and Graph-cut 27
3.2.1 Discriminative Learning based on Linear-SVM 27

3.2.2 Graph-cut based on Spectral Clustering 28
3.3 Experiments 30
3.3.1 Experimental Settings 30
3.3.2 Performance Comparison 32
3.3.3 Model Analysis 35
3.4 Summary 38

4 Representation Learning for Small Object Recognition 40

4.1 Problem Definition 45
4.2 Representation Learning GAN 47
4.2.1 Overview 48
4.2.2 Residual-learning based Generator 49
4.2.3 Adaptive Channel Attention Module 52
4.2.4 Adversarial Learning based Discriminator 53
4.2.5 Feature-attention *RL*-GAN for LR Image Classification 54
4.3 WIDER-SHIP Dataset 55
4.4 Experiments 57
4.4.1 Datasets and Evaluation 57
4.4.2 Implementation Details 58
4.4.3 Performance Comparison 59
4.4.4 Effectiveness of the Residual-learning based G 61
4.4.5 Inputting Features from Lower Layers to G 62
4.4.6 Visualization of the Features Selected by the Feature Attention Module 63
4.5 Summary 63

5 Dual-Stream Representation Learning for Small Object Recognition 65

5.1 Methodology 70

5.1.1 LF-HF Inconsistency Problem 70

5.1.2 Dual-Stream Representation Learning GAN 72

5.1.3 Dual-Channel Image Decomposition 74

5.2 Experiments 77

5.2.1 Datasets and Evaluation 78

5.2.2 Implementation Details 79

5.2.3 Performance Comparison 79

5.2.4 Ablation Study 82

5.3 Summary 85

6 Principal-axis Orientation Distribution for Data Augmentation 86

6.1 Overview of the Proposed System 88

6.2 The Proposed Method 89

6.2.1 Estimating the Orientation of an Object in an Image 90

6.2.2 Similarity Measure based on the Mode 92

6.2.3 Rotating Images with the Permutation Matrix 94

6.3 Experiment Results 95

6.3.1 The Impact of Rotating Testing Images on the Performance of CNNs in Object Detection 96

6.3.2 Estimation of Object Orientation and Comparison with Existing Method 99

6.3.3 Similarity Measure of Principal-axis Orientation Distribution . 101

6.3.4 Data Augmentation Applied to Object Detection 102

6.4 Summary 102

7 Conclusion and Future Work 104

7.1 Conclusion 104

7.1.1 Context-Based Small Object Detection 104

7.1.2 Representation Learning for Small Object Recognition 105

7.1.3 Dual-Stream Representation Learning for Small Object Recognition 105

7.1.4 Principal-axis Orientation Distribution for Data Augmentation 106

7.2 Future Work 106

Bibliography 108

List of Figures

1.1 Typical applications of small object detection and recognition. 2

1.2 An overview of the research presented in this thesis. 8

3.1 Tiny faces detected with our proposed approach (shown as yellow and green boxes) and the HR approach [45] (shown as yellow boxes). 23

3.2 The framework of our proposed DLGC for high-density tiny face detection. 26

3.3 The framework of our proposed DLGC. In the stage for discriminative learning, each candidate pair is shown as boxes with same colours. Some pairs are true matches (top right), while other are false matches (bottom left). The blue or yellow arrows pointing left and right means a max-margin separating hyperplane. 27

3.4 Qualitative results of spectral clustering given similarity matrix between different candidates. Faces (shown as green rectangle) and background regions (shown as red rectangle) are clustered into different groups. 29

3.5 The precision-recall (PR) curves obtained using our proposed DLGC approach and the-state-of-the-arts. 31

3.6 Qualitative results of object detection with our method in RSIs (shown as yellow and green boxes) and Faster RCNN (shown as yellow boxes). Red boxes are introduced by reducing classification threshold and removed by our method. 34

3.7 The PR curves obtained on WIDER FACE validation set using our proposed DLGC approach and the-state-of-the-arts. 39

4.1 HR region of interest (RoI) and LR RoI exhibit different representations from high-level convolutional layers of the CNN classifier. We propose *RL*-GAN to generate the representations from LR RoI to be similar to HR RoI, thus improving the recognition performance on LR objects. 41

4.2 Details of the proposed feature-attention *RL*-GAN. The context enclosed by the blue dotted lines is a standard CNN for object recognition. (a) The residual feature generator is a deep residual network which takes the features from lower-level layers as input and learns the residual feature between the HR and LR images in feature representation. Then, the enhanced representation is achieved through element-wise sum operation between the residual feature and LR representation. (b) the feature discriminator takes the features of the enhanced representation (fake samples) of LR images and the features of the HR images (real samples) as input and tries to differentiate them. (c) the feature attention module selectively emphasises or suppresses features on different channels using global feature. 43

4.3 The architecture of the residual-learning based generator. The input is the feature map $F_i(x_{lr}), 1 \leq i \leq N$, and its output is the residual representation $G(F_i(x_{lr}))$. 51

4.4 The architecture of the adaptive channel attention module. 53

4.5 The architecture of an adversarial-learning based discriminator architecture. It attempts to differentiate between the high-resolution feature representation $F_Q(x_{hr})$ and the regenerated low-resolution feature representation $E_i(x_{lr})$. 54

4.6 Example ship images in the WIDER-SHIP dataset. 56

4.7 Statistics of instances in WIDER-SHIP. 57

4.8 Examples of original and downsampled images in CIFAR-10. Images in the odd rows are of high-resolution (32×32), and images in the even rows are the corresponding low-resolution (8×8) ones. 58

4.9 Visualisation of the input feature map, scaling factors and the rescaled feature map generated by Feature Attention Module at the 2nd Convolution Layer of the backbone. The feature maps containing rich ship information (shown in red solid boxes) have a corresponding higher scale factor, whereas the feature maps containing little ship information (shown in pink broken-line boxes) have a much lower scale factor. 64

5.1 Illustration of the major existing ideas attempting to address the low-resolution image classification problem. (a) is a standard image classifier [40]. (b) is an image-enhancement based idea [103]. (c) is a representation-enhancement based idea [94]. (d) is our proposed simultaneous representation-enhancement idea. 66

5.2 Overview of the proposed *DRL*-GAN. The whole framework is composed of an Image Decomposition module ϕ, Low Frequency and High Frequency Generators G^L and G^H, the corresponding Low Frequency and High Frequency Discriminators D^L and D^H, and a Classifier C. The Image Decomposition Module is used to decompose the input LR and HR images into their low and high frequency components, respectively. G^L and G^H generate enhanced low and high-frequency components of the LR representations by recovering their missing information. D^L and D^H differentiate the enhanced low and high-frequency components of LR representations from their HR counterparts. The enhanced LF (blue rectangle) and HF (red rectangle) representations are concatenated to form the super representation for final classification. In the HR image flow, the low and high-frequency components of the HR images are sent into D^L and D^H as real samples to guide the adversarial learning. In the LR image flow, the low and high-frequency components of the LR images are fed into G^L and G^H respectively to generate their corresponding enhanced representations. The enhanced LF and HF components are also fed into D^L and D^H as fake samples for adversarial learning. 69

5.3 Dual-Channel Auto-encoder for image decomposition. We use an auto-encoder module to decompose an input image into two channels, which carry LF and HF information, respectively, and to reconstruct the input image using the LF and HF Decoders. The black arrows in left sub-figure indicate the process of a standard auto-encoder. 75

5.4 Samples of original images, reconstructed images, and decomposed and reconstructed images in HF and LF channels, respectively. 83

5.5 Visualization of decomposed HR images and the corresponding LR images reconstructed by G. 84

6.1 Illustration of our proposed data augmentation algorithm. 88

6.2 Estimating the orientation of an object based on super-pixel segmentation. 89

6.3 Estimation of super-pixel orientation β_i and coordinate estimation of mode in $H_i(\alpha)$ based on linear interpolation. 92

6.4 The object detection results of the Faster RCNN on VOC2007. The X-axis indicates the rotation angle ranging from 0° to 359° with an interval of 0.5°, and the Y-axis represents the resultant AP. The mAP is obtained overall object categories with the Faster RCNN. . . 97

6.5 The object detection result obtained with the RCNN. The X-axis indicates the rotation angle ranging from 0° to 359° with an interval of the testing set is 0.5°, and the Y-axis represents AP. Again mAP is obtained from all object categories. 98

6.6 The detection mAP and AP obtained with the Faster RCNN over the categories of TV monitor, Cow, Cat, Bird, Dog, Bicycle, Motorbike and Sheep. The X-axis indicates that the rotation angle ranging from 0° to 359° with an interval of 0.5°, and the Y-axis represents the AP. The mAP is again obtained overall categories with the Faster RCNN. 99

6.7 Results of our method for estimating object orientation. (a) −45.13°, (b) 35.75°, (c) 43.13°, (d) −120.35°, (e) −45.68°, (f) −81.39°, (g) −100.27°, (h) −15.17°, (i) −78.08° and (j) 81.15°. The yellow arrows in each image represent the orientation which our method estimated. 100

6.8 The distribution of the principal-axis orientation. 100

6.9 Data augmentation in training set. The X axis indicates that the rotation angle interval is 3° in the range of [0°, 180°]. The Y axis represents the AP and mAP of the Faster RCNN achieved over the categories of "TV monitor". 101

List of Tables

1.1 The author's contribution to SODR in the thesis. 11

3.1 The Average Precision of Classification on the DOTA Validation set . 35

3.2 The Average Precision of Classification on DOTA Testing set 36

3.3 The accuracy of classification based on SVM 36

3.4 The accuracy of classification based on neural network 37

4.1 Comparison of the existing ship datasets 55

4.2 WIDER-SHIP: Comparison of classification accuracy (%) with or without feature-attention *RL*-GAN on different image resolutions (in terms of metres per pixel). 60

4.3 Comparison of classification accuracy (%) on the HRSC dataset with and without using our feature-attention *RL*-GAN on different image resolutions (*i.e.*, in terms of metres per pixel). 60

4.4 Classification error rates obtained on the CIFAR-10 testing set. . . . 61

4.5 Comparisons of classification accuracy (%) of using features from different layers. 63

5.1 WIDER-SHIP: Comparison of classification accuracies (%) of benchmark models with *DRL*-GAN with various image resolutions (in terms of metres per pixel). 80

5.2 Comparison of classification accuracies (%) of benchmark models on the HRSC dataset with our *DRL*-GAN at various image resolutions, *i.e.*, metres per pixel. 81

5.3 Comparison of classification error rates on the CIFAR-10 testing set. 81

5.4 Comparison of the classification accuracies (%) on the WIDER-SHIP using images in the CH^H, $G^H(CH^H)$, CH^L, $G^L(CH^L)$, $CH^H \cup CH^L$, and $G^H(CH^H) \cup G^L(CH^L)$ domains, respectively. Resolution refers to metres per pixel, and *DRL*-GAN represents $G^H(CH^H) \cup G^L(CH^L)$. 85

Abbreviation

SODR - Small Object Detection and Recognition

RL-GAN - Representation Learning Generative Adversarial Learning

DRL-GAN - Dual-stream Representation Learning Generative Adversarial Learning

CNNs - Convolutional neural networks

MLGC - Metric Learning and Graph-cut

DLGC - Discrimative Learning and Graph-cut

PR - Precision Recall

HR - High Resolution

LR - Low Resolution

SR - Super Resolution

RoI - Region of Interest

WIDER-SHIP - A New Data for Ship Recognition

SCS - Small object detection in Crowded Scenes

RSIs - Remote Sensing Images

AP - Averages Precision

mAP - mean of Averages Precision

PODs - Principal-axis Orientations

LF - Low Frequency

HF - High Frequency